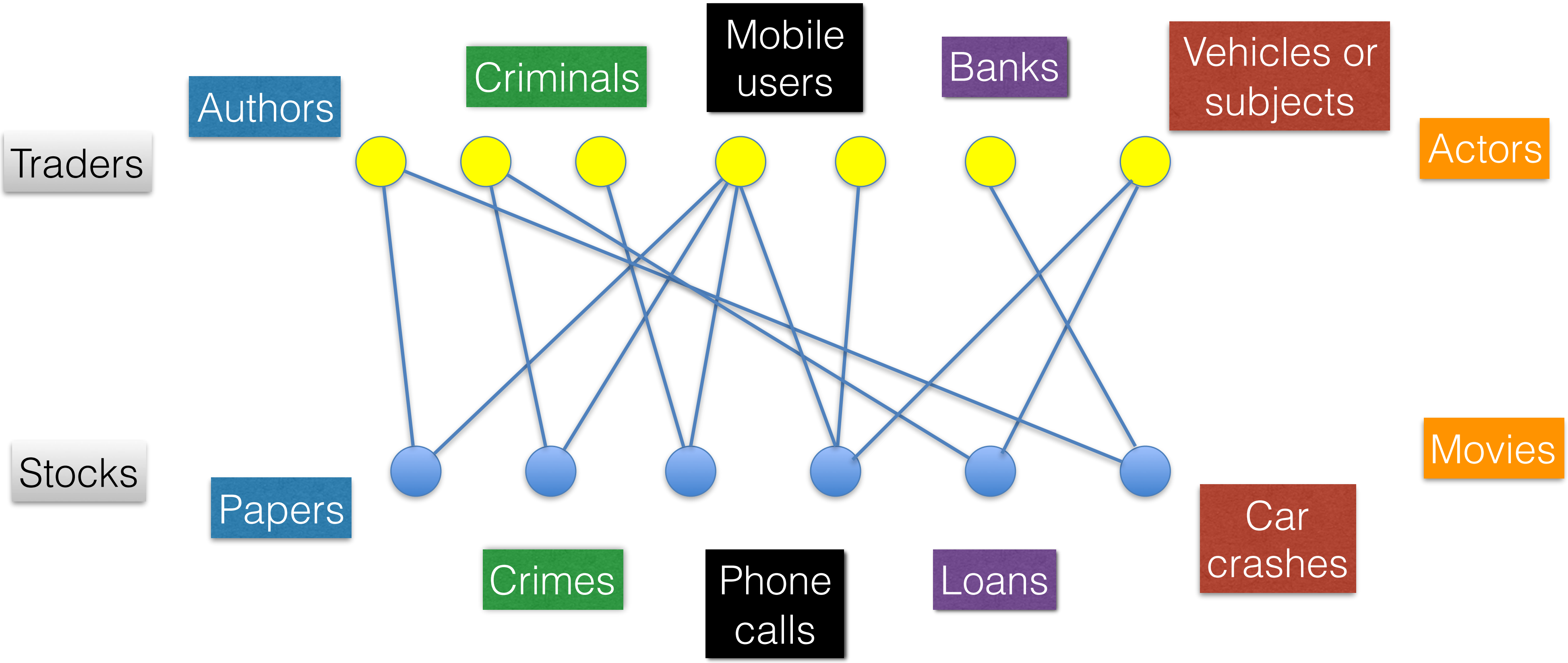
Network analysis: Criminal specialization and fraud detection

Michele Tumminello, Andrea Consiglio
Department of Economics, Management and Statistics
University of Palermo

Summary

- Bipartite Networks and statistically validated networks
- The integrated Antifraud Archive
- Network indicators
- Criminal specialization and network motifs
- Conclusions

Bipartite networks



Traders

Authors

Criminals

Mobile users

Banks

Vehicles or subjects

Actors

Stocks

Papers

Crimes

Phone calls

Loans

Car crashes

Movies

NJP
2012
QF
2015

Plos ONE
2013

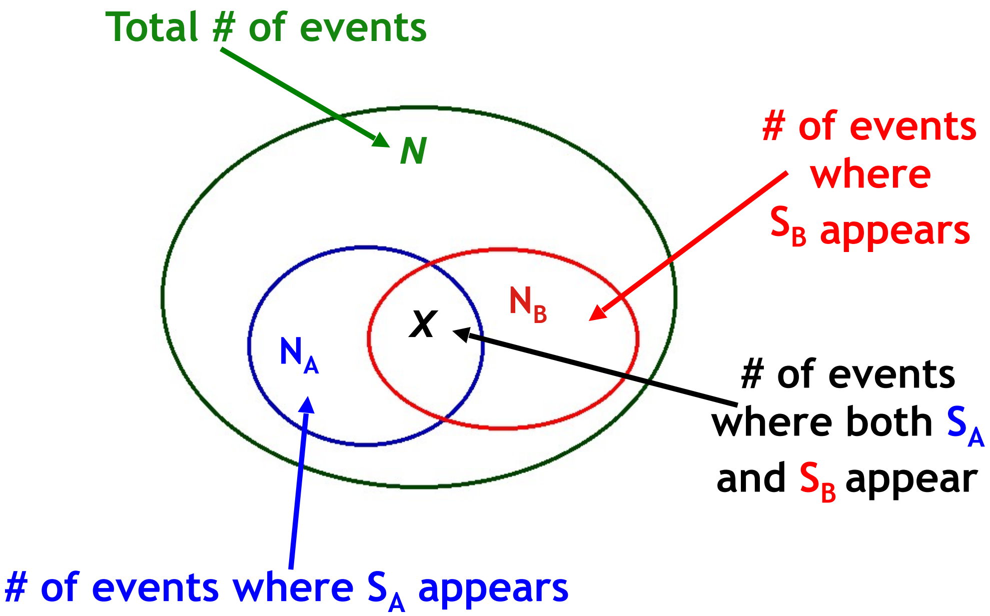
NJP
2015

QF
2015
JEDC
2015

Plos ONE
2011

A statistical validation of co-occurrence

Suppose there are **N** events in the investigated set. Suppose we want to statistically validate the co-occurrence of subject $\mathbf{S_A}$ and subject $\mathbf{S_B}$. Suppose that the number of events where $\mathbf{S_A}$ ($\mathbf{S_B}$) appears is $\mathbf{N_A}$ ($\mathbf{N_B}$), whereas the number of events where both $\mathbf{S_A}$ and $\mathbf{S_B}$ appear is ***X***.


Total # of events
N
# of events where S_B appears
N_B
X
N_A
# of events where both S_A and S_B appear
# of events where S_A appears

The question that characterizes the null hypothesis is:
what is the probability that the number X occurs by chance?

Tumminello M, Miccichè S, Lillo F, Piilo J, Mantegna RN (2011) Statistically Validated Networks in Bipartite Complex Systems. PLOS ONE 6(3): e17994. doi:10.1371/journal.pone.0017994
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017994

Hypergeometric distribution and Statistically Validated Networks

p-value associated with a detection of co-occurrences ≥ X:

$$p = \sum_{i=X}^{Min(N_A, N_B)} \frac{\binom{N_A}{i}\binom{N - N_A}{N_B - i}}{\binom{N}{N_B}}$$

- Count the total number of tests: *T*

- Arrange *p-values* in increasing order.

- Set a link between two vertices if the associated p-value satisfies one of the following inequalities

Bonferroni correction : $p - value_{(k)} < \frac{\alpha}{T}$ → **Bonferroni Network**

Holm-Bonferroni correction : $p - value_{(k)} < \frac{\alpha}{T - k}$ → **Holm-Bonferroni Network**

FDR correction : $p - value_{(k)} < \frac{\alpha\, k}{T}$ → **FDR Network**

Type I error control: false positive links

Proposition 1: the probability that a false positive link is set in the **Bonferroni network** is smaller than α .

Co-occurrences might be dependent

Bonferroni network

- It’s the most conservative network
- The test is data independent
- A **co-occurence** equal to **1** is not statistically significant, provided that the number of links, E, in the co-occurrence network is larger than the number of nodes in the projected set divided by α

$$p{-}value(n_{AB}=1|N_A,N_B,N)=N_A\,N_B\,\frac{(N-N_A)!\,(N-N_B)!}{(N-N_A-N_B+1)!}\geq p{-}value(n_{AB}=1|1,1,N)=\frac{1}{N}>\frac{0.01}{E}$$

Type I error control: false positive links

Proposition 2: the probability that a false positive link is set in the **Holm-Bonferroni network** is smaller than α .

Proposition 3: the expected proportion of false positive links in the **FDR network** is smaller than α, under the (*unrealistic*) assumption that co-occurrences are independent.

The Integrated Antifraud Archive (AIA)

- Time period: 2011-2016
- About 14 million car crashes
- About 20 million individuals and companies
- About 18 million vehicles

Tumminello M, Consiglio A, **Project**: "*Network analysis and modelling of the integrated anti-fraud database*", funded by the Istituto per la Vigilanza sulle Assicurazioni (**IVASS**), which is the National Agency that supervises the activity of all the insurance companies operating in Italy. Responsible for IVASS: **Farabullini F**

Distinguishing between subjects and vehicles

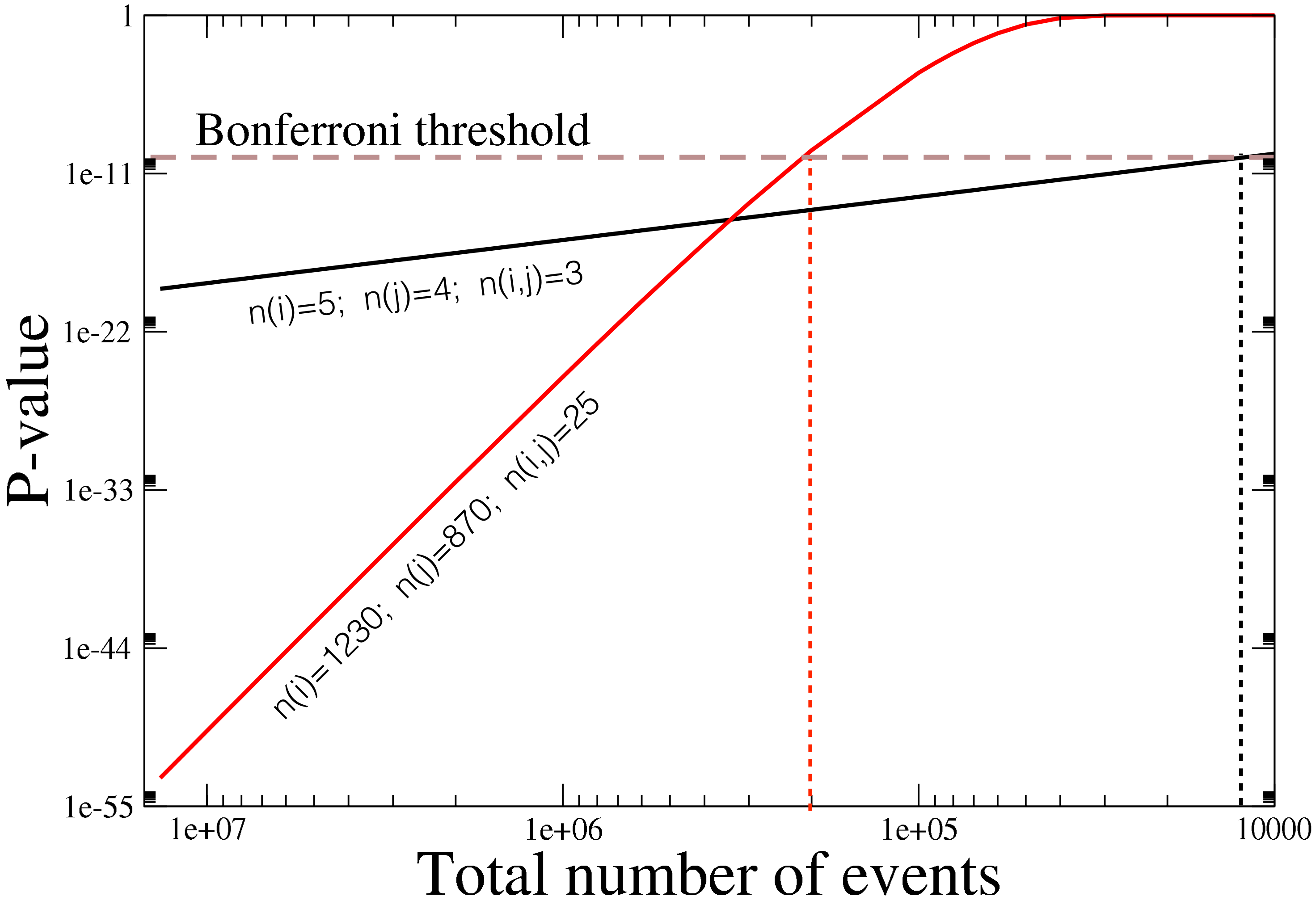
	Nodes	Links	Connected components (CC)	Size of largest CC
Bonferroni network of **subjects***	1,197,055	1,113,389	407.552	**318,876**
Bonferroni network of **vehicles***	209,801	121.253	99,373	**11**

*Subjects and vehicles recorded in the white list have been excluded from the analysis

Bonferroni network: heterogeneity of subjects

Number of events per subject	Subjects in the bipartite network	Difference btw Subjects in contiguous groups	Events in the bipartite network	Subjects in the Bonferroni network	Links in the Bonferroni network	Subjects in the largest connected component
Any	18,877,177	-	13,533,500	1,197,055	1,113,389	318,876
Less than 10,000	18,877,036	**141**	13,518,704	1,195,356	1,074,812	307,436
Less than 5,000	18,876,613	423	13,505,765	1,187,001	1,006,892	279,945
Less than 1,000	18,873,771	2842	13,473,986	1,156,706	826,475	170,671
Less than 500	18,871,669	2102	13,462,713	1,149,780	788,115	130,562
Less than 100	18,856,567	15102	13,437,058	1,101,720	694,210	844

An indicator of link-robustness to geographical localization



An indicator of link-robustness to localization

T=total number of events in the dataset (**T**=13,533,500 in AIA 10/2016)

B=bonferroni threshold in the dataset (**B**=1.356e-10 in AIA 10/2016)

M(i,j)=Min(Q) such that p-value(n(i),n(j),n(i,j),Q)<**B**

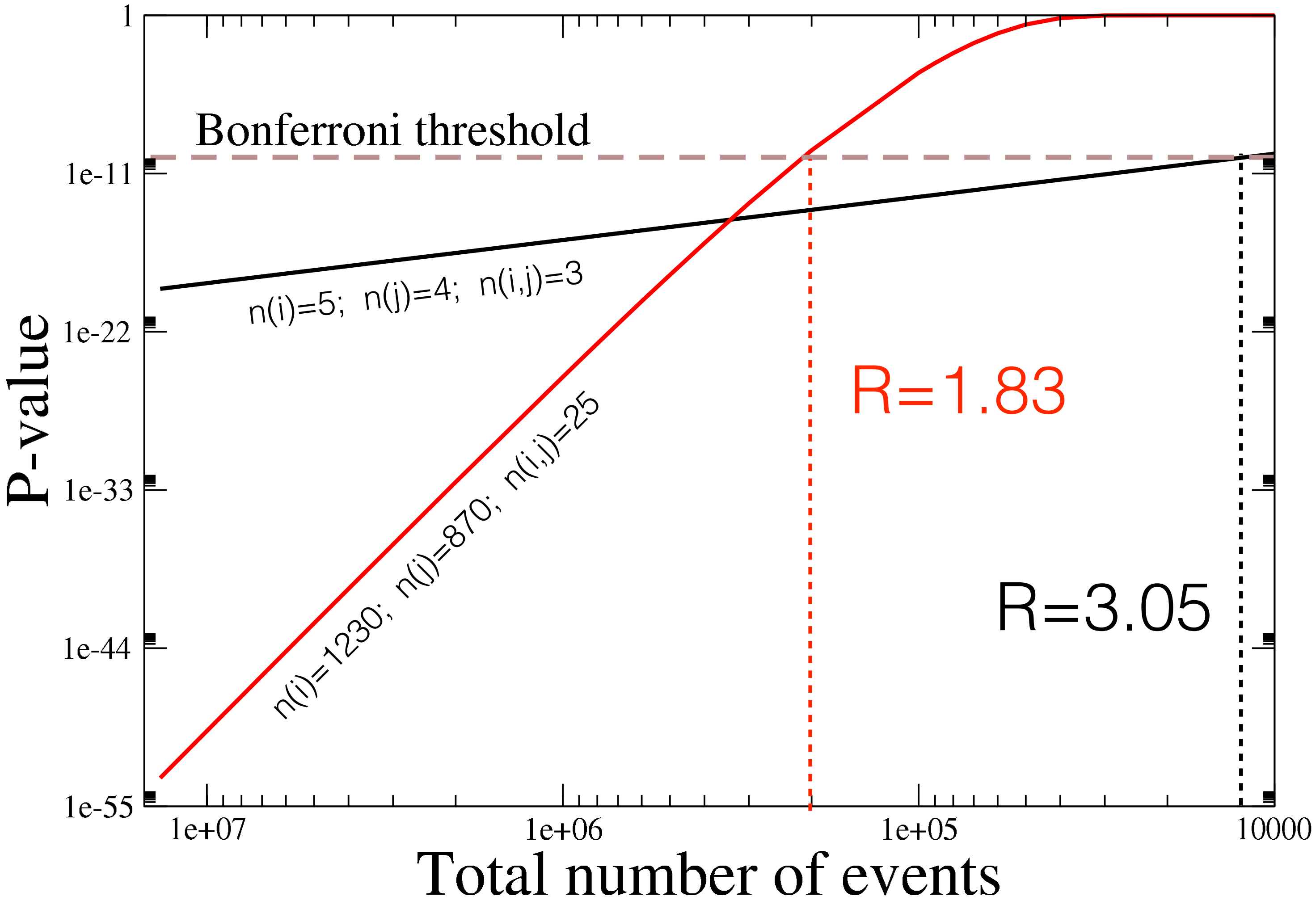
Robustness indicator

$$R(i,j)=log_{10}(T)-log_{10}(M)$$

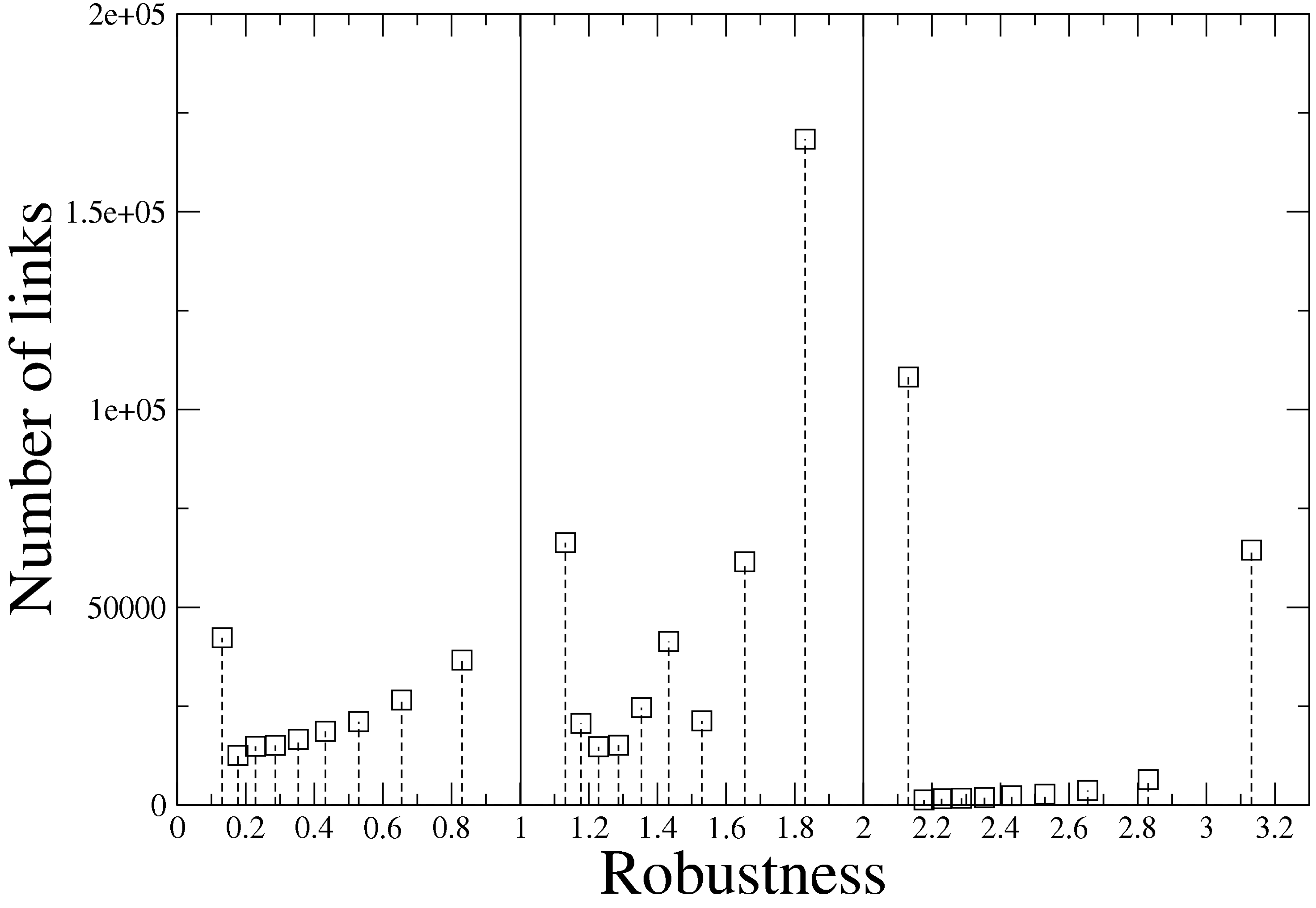
Properties:

- Positivity
- Fast evaluation

An indicator of link-robustness to localization: the rationale



Bonferroni network: distribution of link-robustness



Indicators

- Network level indicators
- Event/subject/vehicle level indicators

Subject indicators

The R indicator is an indicator of link robustness that can be used to construct an indicator of node relevance and/or centrality

Subject strength: $s(i) = \sum_{j=1}^{N(i)} R_{i,j}$

Subject average strength: $as(i) = \frac{\sum_{j=1}^{N(i)} R_{i,j}}{N(i)}$

(relevant, weighted, easy, and fast)

Subject betweenness: $b(i) = \sum_{p,q} \frac{\sigma_{p,q}(i)}{\sigma_{p,q}}$, where $\sigma_{p,q}$ is the number of shortest paths between p and q and $\sigma_{p,q}(i)$ is the number of those passing through i.

(relevant, unweighted, more complicated, slow)

Event indicators

For any event ***e***, the list *L(**e**)* of subject pairs with a validated connection "enhanced" by event ***e*** is compiled.

$$\text{Event strength: } s(e) = \textstyle\sum_{(i,j)\in L(e)} R_{i,j}$$

(meaningful, weighted, easy, and fast)

Event betweenness: theoretically easy, but unfeasible in practice (best guess)

Validated bipartite

VALIDATED BIPARTITE:

Given the SVN of subjects (or vehicles), a bipartite network is reconstructed by

- selecting from the original bipartite network all of the ***event(i)-subject(j)*** pairs such that ***event i*** contributed to a **link in the SVN between *subject(j)*** and (at least) another subject.

- finally adding all the subjects involved in the selected events.

K-H core of a bipartite network

The K-H core of a bipartite network is the largest bipartite **subnetwork** such that nodes of Set A have degree at least K and nodes of set B have degree at least H

Network indicators: Mixed event-subject indicators of centrality: the **K-H core**

Event oriented event-subject indicator:

$$KH_e(e,s) = \max(K) \text{ such that } (e,s) \in K - H \text{ core}$$

Subject oriented event-subject indicator:

$$KH_s(e,s) = \max(H) \text{ such that } (e,s) \in K - H \text{ core}$$

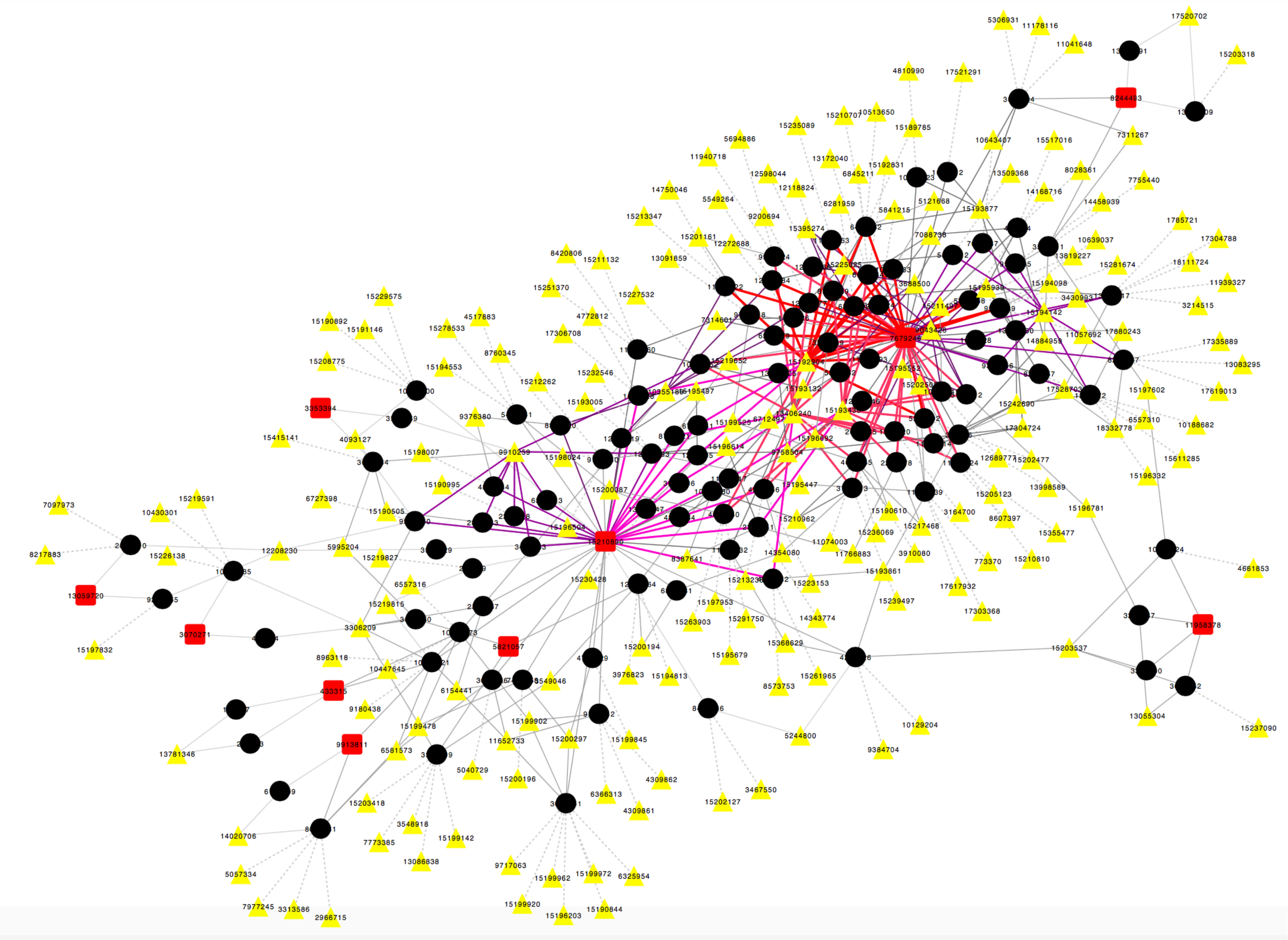
Balanced event-subject indicator:

$$KH(e,s) = \max(\sqrt{K \cdot H}) \text{ such that } (e,s) \in K - H \text{ core}$$

K-H CORE DECOMPOSITION

of a validated bipartite community

(with K>1 and H>1)



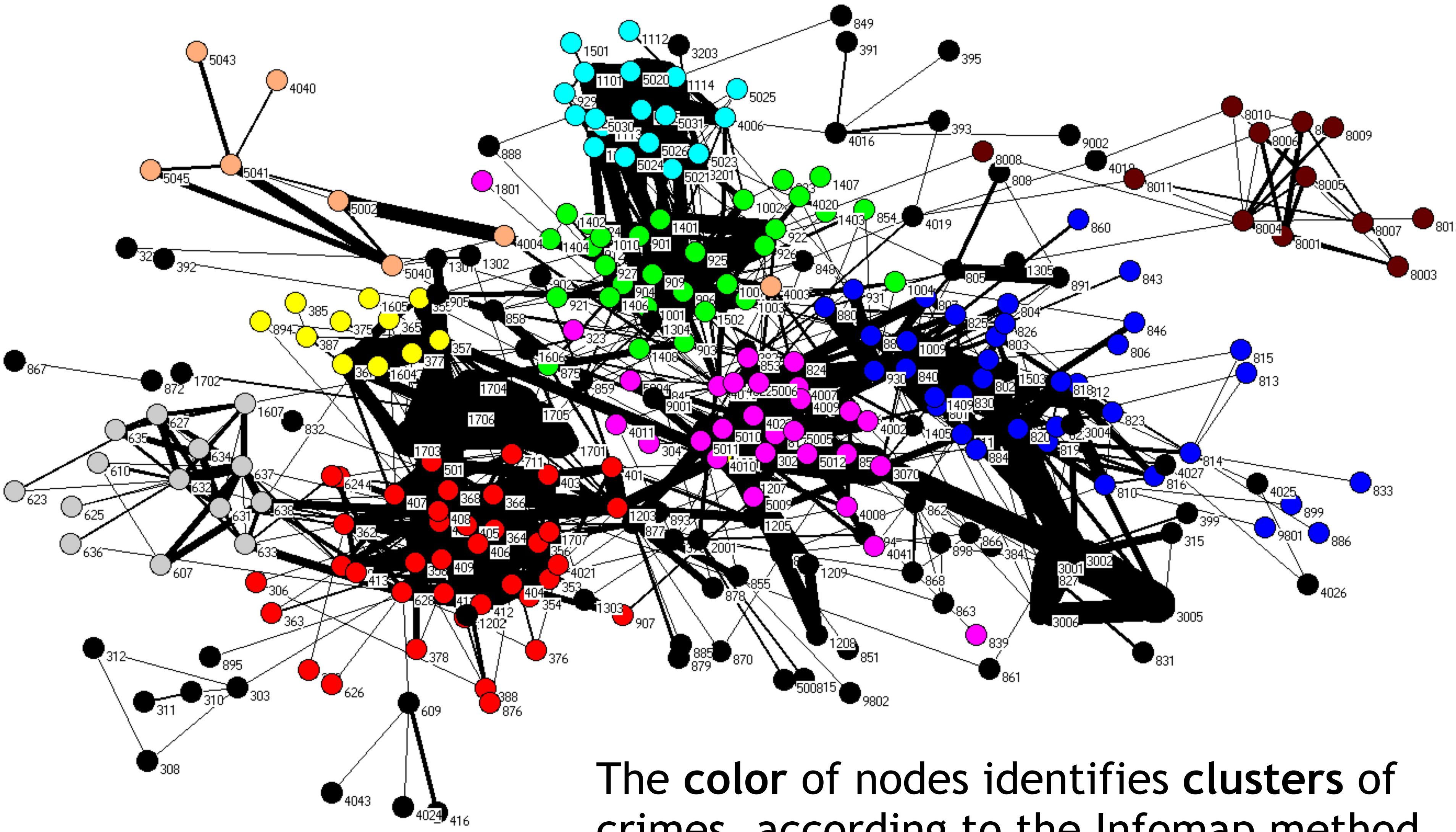
Interlude: criminal specialization

The network of crimes

- We have a list of 336,069 individuals who have been suspected of at least one crime over one decade time window: about 2,000,000 instances.
- Crimes are coded in a list of 376 specific crime types (penal code)
- We have information about gender and age of individuals.

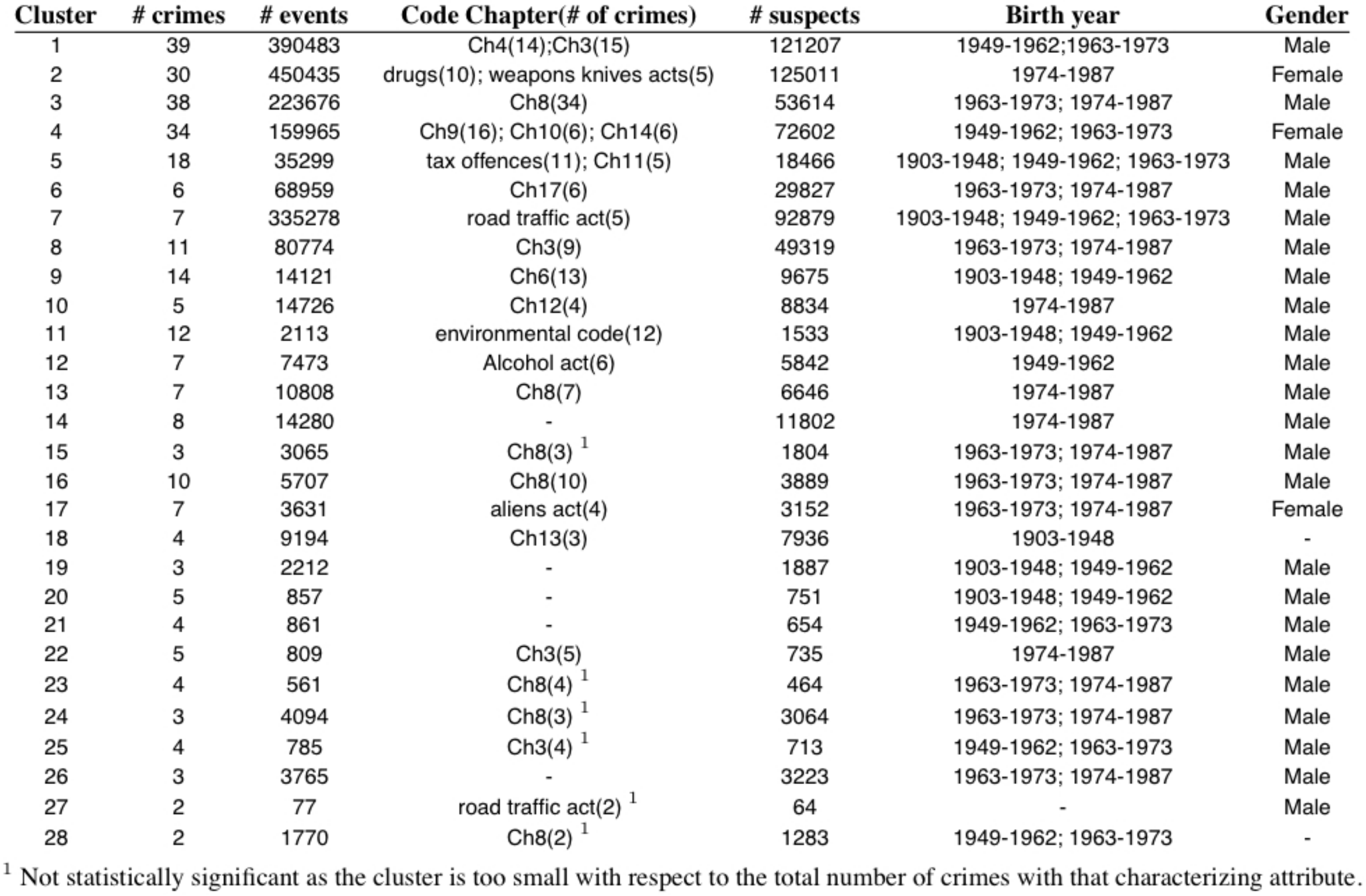
M Tumminello, C Edling, F Liljeros, RN Mantegna, J Sarnecki (2013) The Phenomenology of Specialization of Criminal Suspects. PLoS ONE 8(5): e64703. doi:10.1371/journal.pone.0064703

The weighted FDR network of crimes



The **color** of nodes identifies **clusters** of crimes, according to the Infomap method

Characterization of clusters

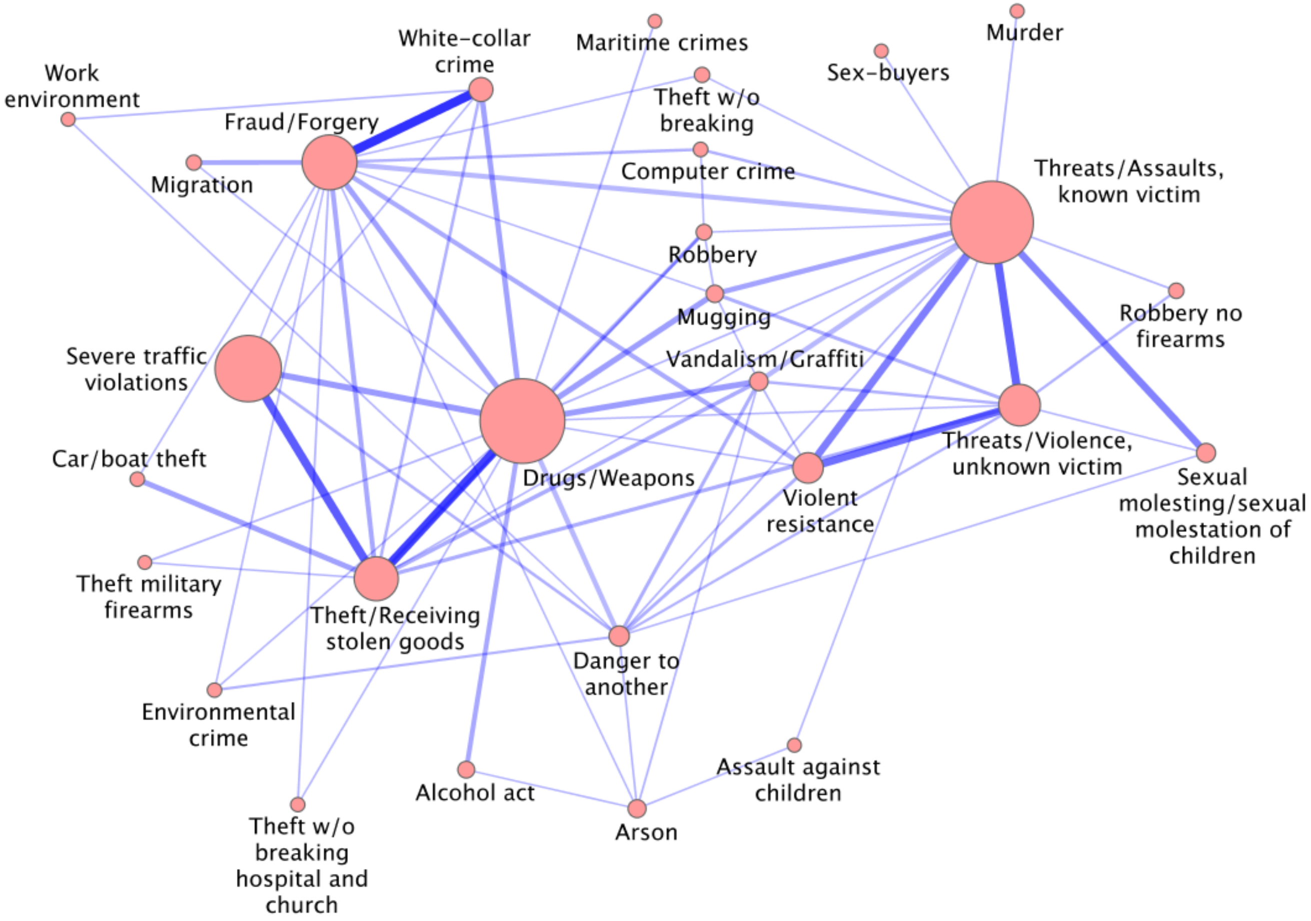


Cluster	# crimes	# events	Code Chapter(# of crimes)	# suspects	Birth year	Gender
1	39	390483	Ch4(14);Ch3(15)	121207	1949-1962;1963-1973	Male
2	30	450435	drugs(10); weapons knives acts(5)	125011	1974-1987	Female
3	38	223676	Ch8(34)	53614	1963-1973; 1974-1987	Male
4	34	159965	Ch9(16); Ch10(6); Ch14(6)	72602	1949-1962; 1963-1973	Female
5	18	35299	tax offences(11); Ch11(5)	18466	1903-1948; 1949-1962; 1963-1973	Male
6	6	68959	Ch17(6)	29827	1963-1973; 1974-1987	Male
7	7	335278	road traffic act(5)	92879	1903-1948; 1949-1962; 1963-1973	Male
8	11	80774	Ch3(9)	49319	1963-1973; 1974-1987	Male
9	14	14121	Ch6(13)	9675	1903-1948; 1949-1962	Male
10	5	14726	Ch12(4)	8834	1974-1987	Male
11	12	2113	environmental code(12)	1533	1903-1948; 1949-1962	Male
12	7	7473	Alcohol act(6)	5842	1949-1962	Male
13	7	10808	Ch8(7)	6646	1974-1987	Male
14	8	14280	-	11802	1974-1987	Male
15	3	3065	Ch8(3) [1]	1804	1963-1973; 1974-1987	Male
16	10	5707	Ch8(10)	3889	1963-1973; 1974-1987	Male
17	7	3631	aliens act(4)	3152	1963-1973; 1974-1987	Female
18	4	9194	Ch13(3)	7936	1903-1948	-
19	3	2212	-	1887	1903-1948; 1949-1962	Male
20	5	857	-	751	1903-1948; 1949-1962	Male
21	4	861	-	654	1949-1962; 1963-1973	Male
22	5	809	Ch3(5)	735	1974-1987	Male
23	4	561	Ch8(4) [1]	464	1963-1973; 1974-1987	Male
24	3	4094	Ch8(3) [1]	3064	1963-1973; 1974-1987	Male
25	4	785	Ch3(4) [1]	713	1949-1962; 1963-1973	Male
26	3	3765	-	3223	1963-1973; 1974-1987	Male
27	2	77	road traffic act(2) [1]	64	-	Male
28	2	1770	Ch8(2) [1]	1283	1949-1962; 1963-1973	-

[1] Not statistically significant as the cluster is too small with respect to the total number of crimes with that characterizing attribute.

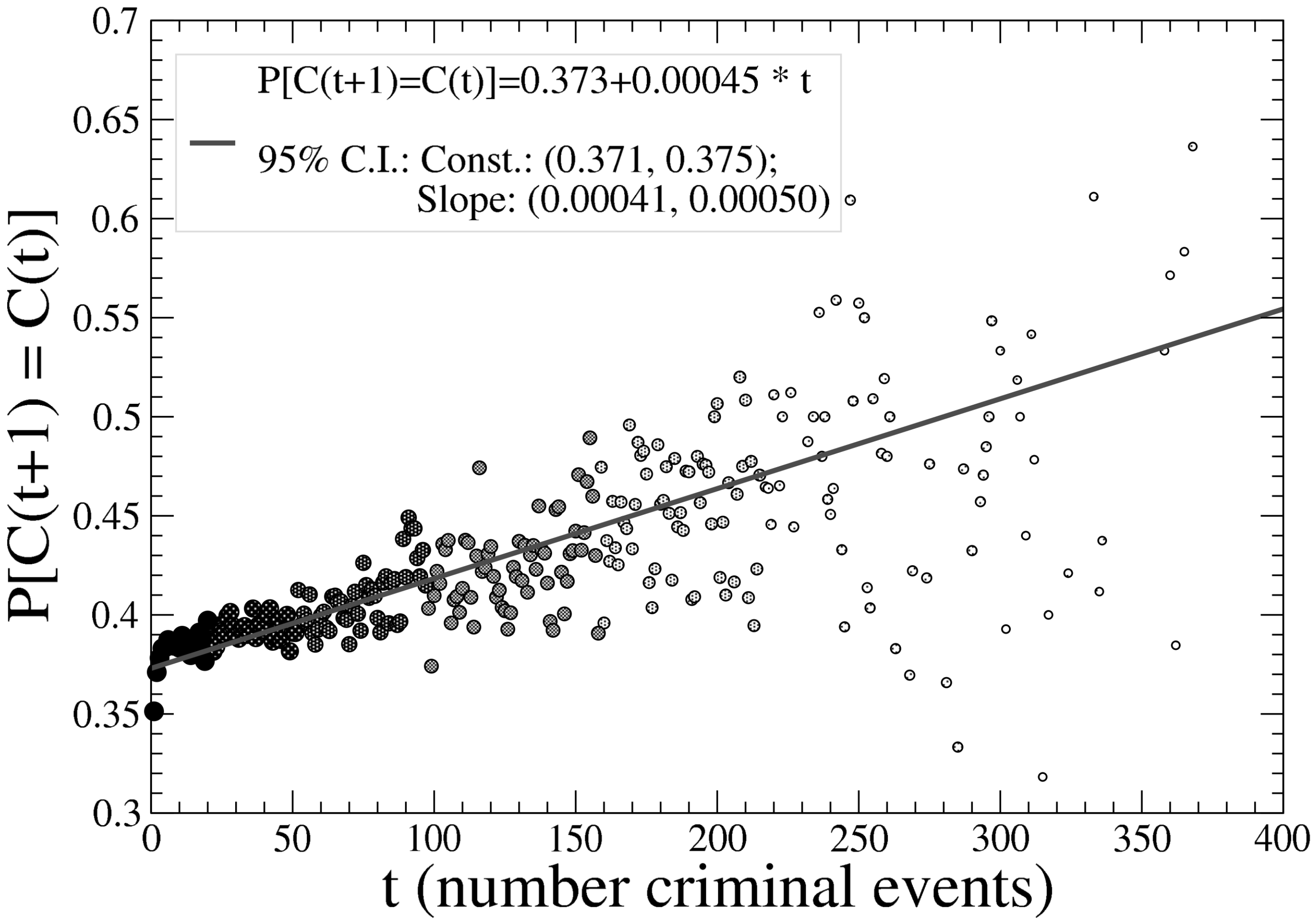
Chapter 3 (Assault) - chapter 4 (Crimes against liberty and peace) – chapter 6 (sexual offences) - chapter 8 (Theft & Robbery) – chapter 9 (Fraud and other acts of dishonesty) – chapter 11 (tax offences) – chapter 12 (environmental offences).

Interpretation of clusters in the FDR network



The method of cluster characterization has been introduced by MT et al. (2011), Community characterization of heterogeneous complex systems, J. Stat. Mech. P01019

Probability that a suspect who has been already suspected of "t" crimes in her career is then suspected of a crime, the "t+1" crime, which belongs to the same cluster as crime "t", as a function of (the proxy of) career progression "t".



"The little specialization, which still exists, occurs after adolescence and increases with criminal career progression"

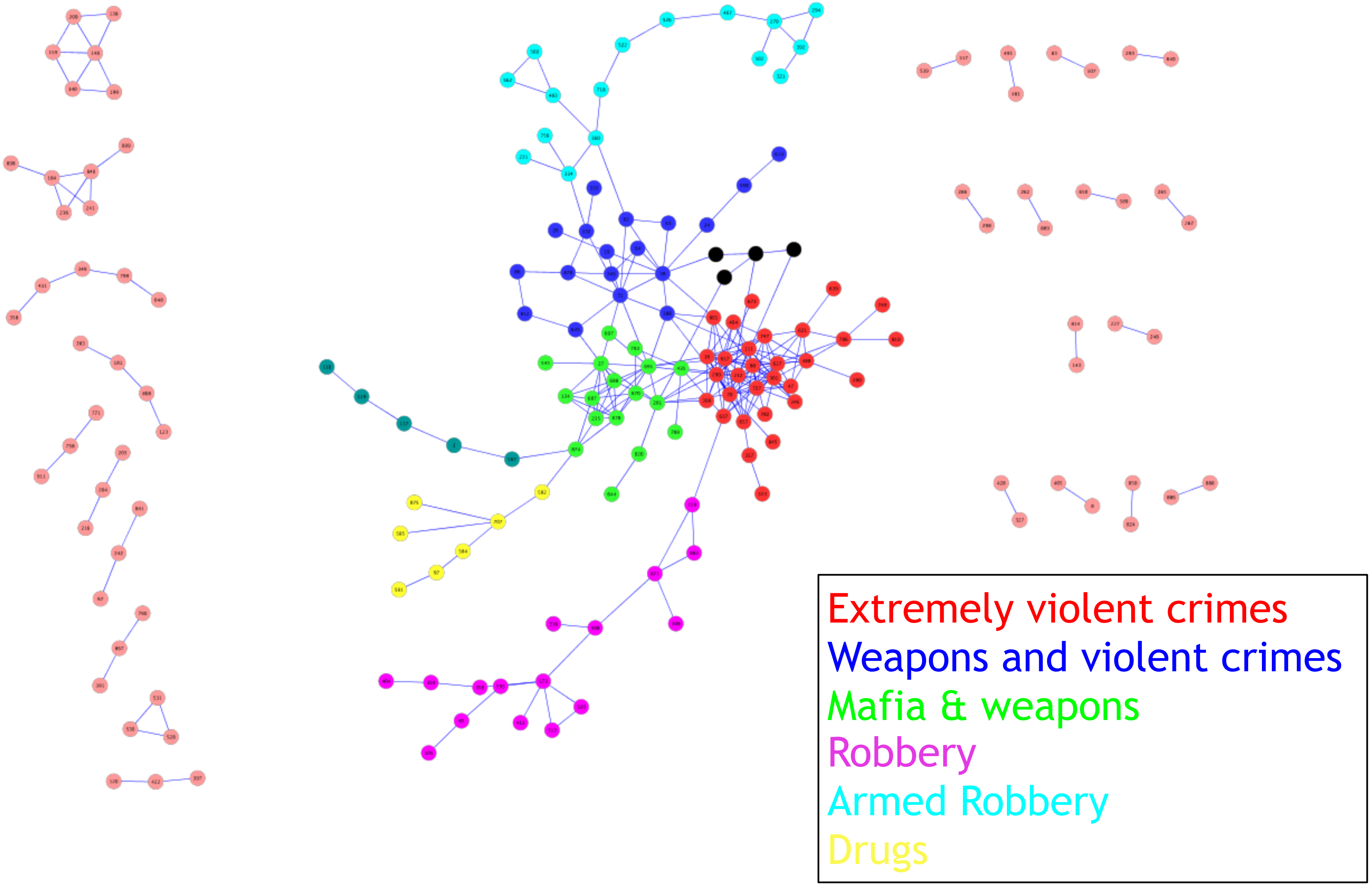
(Blumstein1986,Piquero1999).

Criminal specialization and organized crime

- A **collaboration** between **Procura di Palermo** (Gery Ferrara) and **University of Palermo** (Michele Tumminello and Salvatore Micciche’).

- **Data**:
  - Criminal records (“**casellario giudiziario**”)
  - Detailed vital statistics (“**anagrafica di secondo livello**” – incomplete)

FDR network of crimes



Extremely violent crimes

Weapons and violent crimes

Mafia & weapons

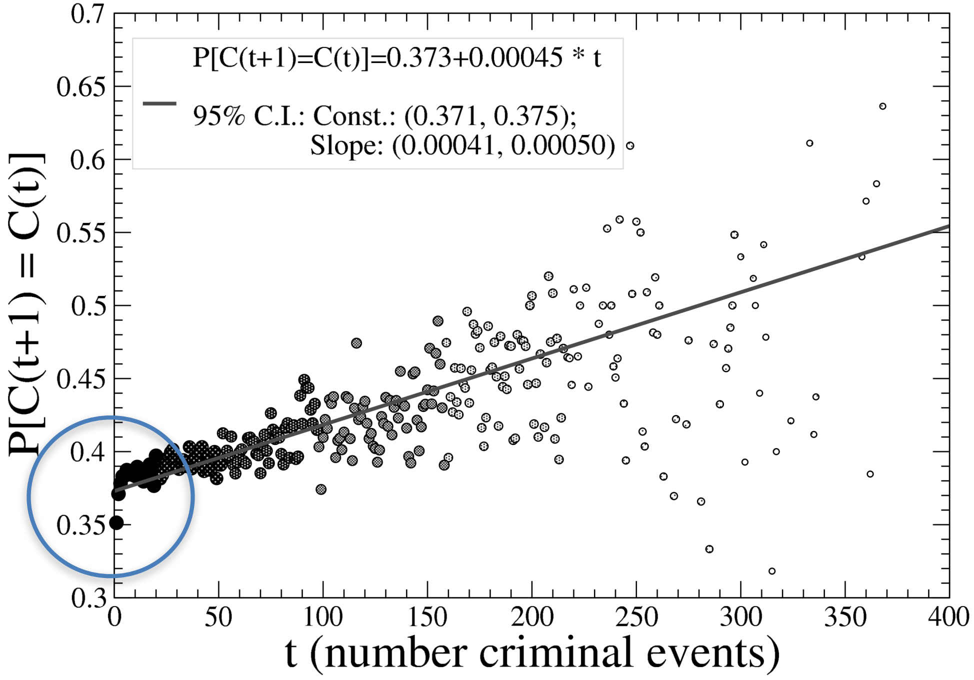
Robbery

Armed Robbery

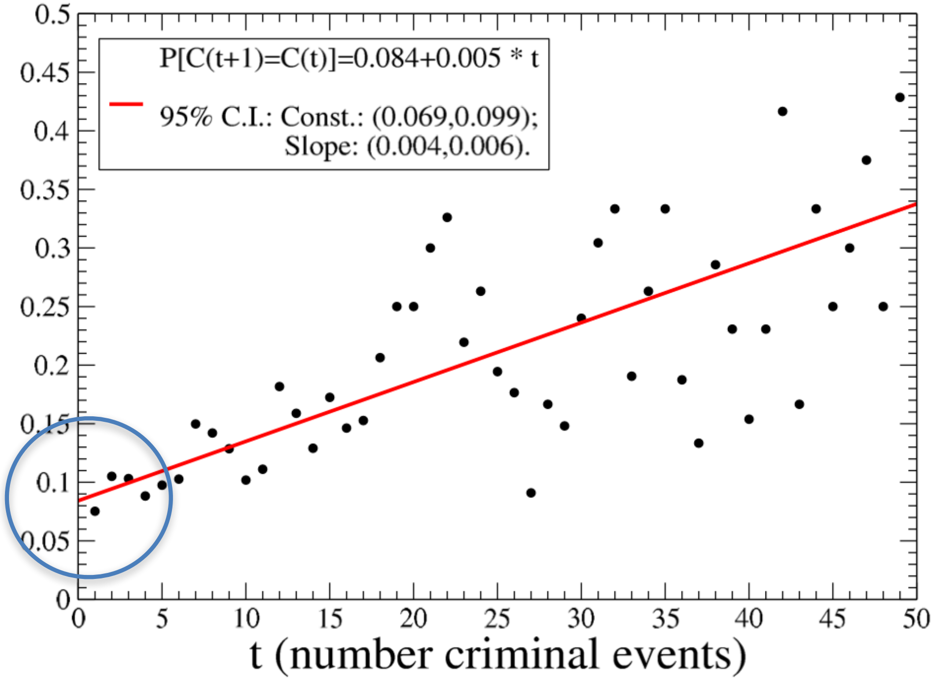
Drugs

Specialization and criminal career

Sweden

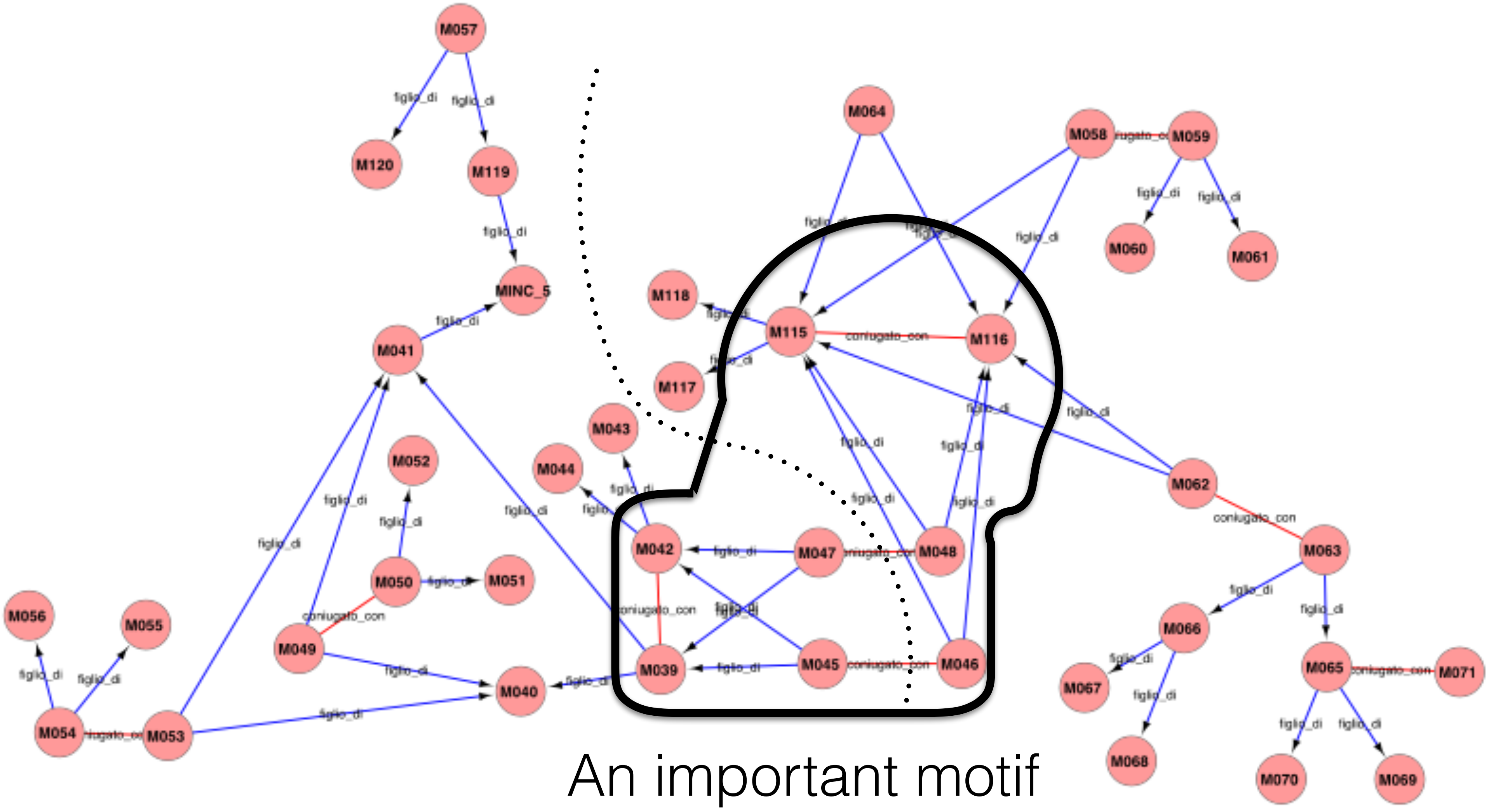


Palermo dataset



At the beginning of their career, criminals included in the Palermo dataset are generalists.

A network of two families



An important motif

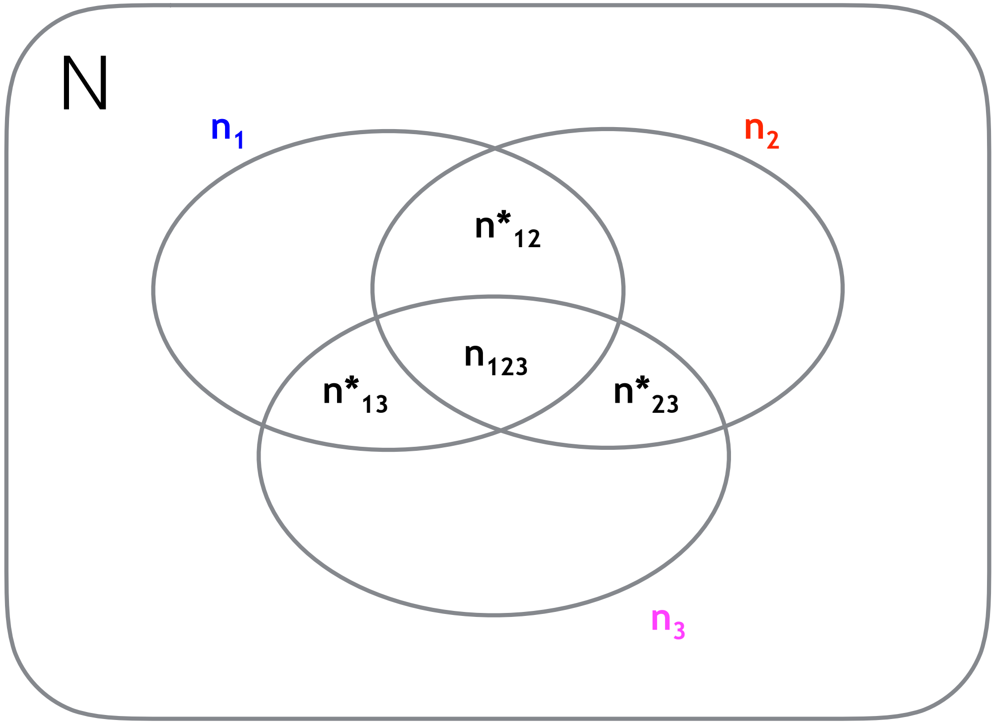
In summary

- Criminal specialization
- Some types of crimes require cooperation
- Cooperation requires coordination

⇩

Motifs

Three-node motifs: triangles



$$p(n^*_{12}, n^*_{13}, n^*_{23} | n_1, n_2, n_3, N) = \sum_{n_{12}} \frac{\binom{n_1}{n_{12}}\binom{N-n_1}{n_2-n_{12}}\binom{n_{12}}{n_{12}-n^*_{12}}\binom{n_1-n_{12}}{n^*_{13}}\binom{n_2-n_{12}}{n^*_{23}}\binom{N-n_1-n_2+n_{12}}{n_3-n^*_{13}-n^*_{23}-n_{12}-n^*_{12}}}{\binom{N}{n_2}\binom{N}{n_3}}$$

$$\text{p-value} = p(n^*_{12} + n^*_{13} + n^*_{23} \geq n^{*,0}_{12} + n^{*,0}_{13} + n^{*,0}_{23})$$

Three-node motifs and antifraud

Network of directly involved subjects (no professionals)

- Number of triangles: 162,409
- Number of statistically validated triangles:60,523

Randomly rewired network of directly involved subjects

- Average number of triangles: 18,535
- Average Number of statistically validated triangles: 0.08

Preliminary conclusions

1. The network of subjects and vehicles carry different information.

2. Considered network indicators and AIA (node) indicators carry complementary information, and, therefore, can fruitfully be integrated.

4. The test on “claims closed following investigation” and the analysis of a few case studies indicate the effectiveness of the overall approach: next step is developing and tuning network indicators with respect to such benchmarks.

Thanks!

Michele Tumminello

Email: michele.tumminello@unipa.it

Alt. Email: michele.tumminello@gmail.com